Project proposal

Project Title: Stroke Prediction and Prevention

Team Members: Diemeng Hu, Julia Wood, Chloe Zhang, Griffin Ong

Background

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. It is necessary to establish a good model and find the important factors affecting stroke so as to achieve the role of prevention and research.

Goal

For our research, we will develop a classifier that predicts the probability of whether or not someone will have a stroke based on relevant input parameters. Healthcare practitioners will be able to identify which patients are at high-risk for a stroke, prescribe proper treatment, and save lives. Such prevention can help save hospitals millions of dollars in stroke-recovery rehabilitation and long-term care. For our business case, we will attempt to quantify the expected value savings of implementing our model for such healthcare institutions.

Description of Dataset

The Stroke Prediction Dataset is derived from Kaggle. This dataset contains 5110 patient information, which includes 4,861 patients with target = 0(no stroke) and 249 patients with target =1(stroke). In addition, the dataset included 10 input parameters to each patient, which may allow us to take into account in the model for prediction of a patient's likelihood of stroke. Among them, categorical variables include gender, hypertension, history of heart disease, marital status, type of work, type of residence, and history of smoking. Numerical variables include age, average glucose level in blood and body mass index.

Methods and Implementation

This project mainly uses supervised learning algorithms to analyze and predict data. We will train the dataset to various models, including but not limited to logistic regression, Bagging, KNN, random forest, SVM, and Neural Networks (ANN). By comparing the accuracy, AUC, RMSE and other relevant parameters of these algorithms, we will get a best and most applicable model which can fit the data very well.

Meanwhile, we will rank the importance of each predictor by using <*varImp*> function in <*caret*> package in R to track the changes in model statistics for each predictor. This step is crucial since predictors that has large impact on models might be considered the key factor for stroke. Hospitals will focus on these factors in patients to prevent stroke.

Comments and Concerns

In this health issue, our priority is focusing on identifying patients who have suffered a stroke. Thus, we would like to confirm any observations that might have stocked. In addition, we'll have to balance our data set before we perform statistical analysis.

We also have concerns about the size of the dataset. 5,110 rows without eliminating missing values in total are relatively small. This may affect we split the data into a sufficient amount of test, training, and validation datasets to affect the accuracy of model performance evaluation.